Polymer sequencing via unsupervised learning of pyrolysis-mass spectra

Authors: Yusuke Hibi*, Shiho Uesaka and Naito Masanobu*

Affiliations:

1Data-driven Polymer Design Group, Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science; 1-2-1, Sengen, Tsukuba, Ibaraki 305-0047, Japan.

*Corresponding authors. Emails: hibi.yusuke@nims.go.jp and naito.masanobu@nims.go.jp

Abstract: A sequence—an arrangement of monomers—dominates polymer properties, as best exemplified by proteins; however, an efficient sequencing method for synthetic polymers has not been established yet. Herein, we propose a polymer sequencer based on mass spectrometry of pyrolyzed oligomeric fragments. By interpreting an observed fragment pattern as one generated from a mixture of sequence-defined copolymers, sequencing can be simplified to compositional analysis. Our key development is a reference-free quantitative mass spectrometry. The reference spectra of the hardly synthesizable sequence-defined copolymers were not measured but virtually inferred via unsupervised learning of the spectral dataset of easily synthesizable random copolymers. The polymer sequencer quantitatively evaluates complex sequence distribution in versatile multi-monomer systems, which would allow sequence–property correlation studies and practical sequence-controlled polymerization.

One-Sentence Summary: Spectral deconvolution into basis spectra extracted from a dataset of random copolymers unveiled sequence distribution.
Main Text:

The impact of a sequence on polymer properties is evident in proteins: the unique sequence defined by RNA determines the three-dimensional structure that holds functional monomers in place for cooperative functioning. However, in vinyl copolymers—the most commonly used synthetic polymers (1)—the sequence–property correlation is less clear owing to the broad sequence distribution inherently and inevitably caused by stochastic copolymerization (2). The copolymer properties are averaged over the distribution; thus, outstanding sequence-originating functions, e.g., self-healing (3), shape-memory (4) and biomolecular recognition (5), can rarely be discovered. To maximize the potential of copolymers, quantitatively evaluating the sequence–property correlation and intentionally biasing the sequence distribution toward the desired direction from the stochastic ground-state predicted by the Lewis-Mayo equation (6), i.e., sequence controlled polymerization, would be crucial. Throughout this process, sequence distribution analysis (7, 8), i.e., sequencing (also see Supplementary Text), plays a fundamental role. Notably, real-time sequencing of growing copolymers could lead to practical sequence-controlled polymerization, because the sequence distribution could be controlled by autonomously optimizing the polymerization conditions using reinforced learning (9) based on the perception of the current sequence status.

Despite these demands, vinyl-copolymer sequencing remains challenging. To analyze sequence distribution, quantification of short sequences (polyads) is necessary, which is difficult for inherently non-quantitative mass spectrometry (MS) utilized in protein sequencing (10). Here, “inherently non-quantitative” means that the abundance ratio of two components cannot be directly determined from the intensity ratio of the corresponding peaks. Furthermore, vinyl-copolymer main chains are chemically robust, limiting the adaptation of tandem-MS sequencing (7), which relies on ionic monomer-to-oligomer fragmentation. Instead of such destructive and non-quantitative MS, non-destructive and quantitative nuclear magnetic resonance spectroscopy (NMR) has been alternatively employed for vinyl-copolymer sequencing (11). The electronic environment of the central monomer unit varies with the adjacent monomer-unit species, inducing peak shifts. However, such shifts are often subtle and coupled with tacticity (12); therefore, the application of NMR is limited to the triad sequencing of specific binary-monomer combinations (7).

To analyze highly complex sequences in versatile multi-monomer systems, we propose a destructive yet quantitative polymer sequencer through ambient gas-phase MS (13) of pyrolyzed oligomeric fragments, wherein the chemically robust main chains undergo thermal decomposition to generate polyads of various lengths/sequences that are recorded as a complicated fragment pattern (Fig. 1A; an ensemble with alternating tendency is illustrated). The peak-intensity ratios potentially reflect the original sequence distribution. However, direct peak-wise characterization is ineffective, because the peak intensities are strongly biased owing to different ionization efficiencies and regioselective thermal cleavages along the chemically inequivalent main chains. Peak-wise calculation of such correction factors is unrealistic. Furthermore, different polyads with identical masses, such as XXY/XYX (X and Y represent two monomer species), are indivisible. To circumvent these difficulties, we interpret the polyad ensemble as one generated from a certain mixture of ideal sequence-defined K-copolymers (basis copolymers) consisting of single-polyad species (Fig. 1A, here K=5). This interpretation is equivalent to assuming that any observed spectrum of the X/Y-copolymers can be approximated to a linear combination of K–basis spectra \( P \in \mathbb{R}_+^{K \times D} \) (\( \mathbb{R}_+^{K \times D} \): non-negative \( K \times D \) matrix; \( D \): channel number). This seems plausible, as polyad distribution primarily depends on the “local”
sequence; the quantity of fragments shorter than the basis polyads can be written in a chain structure (Fig. S1). The basis polyad length is determined by \( K \); e.g., all binary triads and pentads can be expressed using five and nine basis copolymers, respectively (Fig. S1). Peak-wise correction factors are implicitly embedded into the basis spectra. Furthermore, indivisible polyad peaks, e.g., XXY/XYX, can be quantitatively unmixed and attributed to the corresponding basis spectra—in this case, \((XXY)_l\) and \((XY)_l\) (\(l\): a large integer representing repetition). Thus, the sequencing can be simplified to compositional analysis, wherein the fraction \((c_1, \ldots, c_K)\) represents the original sequence distribution (Fig. 1A). The critical challenge now becomes how to prepare the basis-spectral set \( P \). Since there are no general synthetic protocols for sequence-defined copolymerization (2), the basis copolymers, except homopolymers, are unpreparable and, thus, unmeasurable. Our approach to this end is the basis inference via unsupervised learning of spectral dataset \( X \in \mathbb{R}^{N \times D}_+ \) \((N:\text{sample number})\) of easily synthesizable random copolymers, whose polyad fractions are modulated by changing the monomer feed ratios (Fig. 1B). Because random copolymers are also interpretable as mixtures of the basis copolymers (Fig. 1C), every \( n^{th} \) sample spectrum \((X_n)\), stored at the \( n^{th} \) row of \( X \), can be approximated as a linear combination of \( P \) as follows:

\[
X_n \approx \sum_{k=1}^{K} C_{nk} P_k, \quad \text{s.t.} \quad \sum_{k=1}^{K} C_{nk} = 1, (n = 1, \ldots, N)
\]

where \( C \in \mathbb{R}^{N \times K}_+ \) represents the \( K \)-polymer fractions in every sample. This can be simply written as \( X \approx CP \) in the matrix form, which is called non-negative matrix factorization (NMF) (14–16). In this context, NMF is a mathematical expression of compositional analysis that simultaneously identifies bases \( P \) and their quantities \( C \), corresponding to “reference-free” quantitative MS (RQMS). Our NMF-based RQMS-algorithm output reasonable \( C \) and \( P \) (Fig. 1C); the modulation of sequence distribution via the monomer feed ratio was well represented by \( C \), and the triad peaks were rationally unmixed and attributed to five–basis spectra \( P \). After the inference of \( P \), the spectral deconvolution of target spectrum \( X_t \) yields fraction \((c_1, \ldots, c_K)\) as a sequence indicator (Fig. 1B). Herein, we theoretically construct RQMS algorithm based on two-step NMFs, verify its accuracy via benchmark compositional analyses, and apply RQMS to sequencing.
Fig. 1. Vinyl-copolymer sequencing via RQMS. (A) Sequencing simplified to compositional analysis. Fraction \((c_1, \ldots, c_K)\) represents the sequence tendency (here, \(K = 5\)). (B) Flowchart of RQMS sequencing. (C) Spectral dataset \(X \in \mathbb{R}_+^{N \times D}\) of random copolymers (here, \(X/Y = \text{butyl acrylate/styrene}, N = 2I, \text{see the sample information attached to Data S1}\) and factorized \(C \in \mathbb{R}_+^{N \times K}\) and \(P \in \mathbb{R}_+^{K \times D}\).
As there is no established sequencing method generally applicable to versatile monomers, even verification of RQMS sequencing is challenging. We first designed a benchmark system for verifying RQMS accuracy: conducting compositional analysis for ternary films of poly(ethyl methacrylate) (E), poly(methyl methacrylate) (M), and polystyrene (S). The dataset (Data S2) consisted of 24 binary and ternary mixtures, whose highest fractions did not exceed 80 wt% (reference-free). If RQMS-algorithm outputs C and P well consistent to the true composition and the observed pure E/M/S spectra, RQMS sequencing is also promising. This was actually achieved, as shown in Fig. 2A. Even a biased dataset without samples beyond 40 wt% of the S-fraction, gave an accurate estimation (fig. S2). However, direct and single-step polymer-based NMF, \( X \approx CP \), yielded inaccurate results (fig. S3). This is because pyrolysis-MS does not measure polymers themselves, but measure their pyrolyzed fragments; therefore, the spectrally distinct and independent components are not K-polymers but their pyrolyzed M-fragments (\( K \ll M \)). To conduct fragment-based NMF, we assume a latent hierarchical structure (Fig. 2B), where the \( k^{th} \) basis polymer per unit weight generates the \( m^{th} \) fragment spectrum, \( S_m \), with a fragment abundance (FA) of \( B_{km} \) \( (k = 1, ..., K, m = 1, ..., M) \). Here, \( S \in \mathbb{R}_+^{M \times D} \) represents the \( \ell_1 \)-normalized M-fragment spectra, and \( B \in \mathbb{R}_+^{K \times M} \) represents FA of the basis polymers. The observed spectrum of the \( n^{th} \) sample then can be written as a linear combination of \( S \):

\[
X_n \approx \sum_{m=1}^{M} \left( \sum_{k=1}^{K} C_{nk} B_{km} \right) S_m \approx A_n S,
\]

where \( A \approx CB \in \mathbb{R}_+^{N \times M} \) represents the sample-wise FA. The RQMS algorithm first conducted fragment-based NMF, \( X \approx AS \), and subsequently second NMF, \( A \approx CB \), summarized as \( X \approx AS \approx CBS = CP \), where \( BS \in \mathbb{R}_+^{K \times D} \) represents the pure K-polymer spectra equivalent to \( P \). The actual implementation is slightly more complicated, as depicted in figs. S4–5.

Because both NMFs are low-rank approximations, the fundamental driving force toward the optimum solution is identical: minimizing the square residuals, e.g., \( \min_{A \geq 0, S \geq 0} \| X - AS \|_F^2 \). A solution \( (A^*, S^*) \) derived from this criterion is, however, not unique even for a given component number \( M \), as any non-singular \( Q \in \mathbb{R}^{M \times M} \) satisfying \( A^* Q \geq 0, Q^{-1} S^* \geq 0 \) gives another solution \( (A^* Q, Q^{-1} S^*) \) (17). Imposing additional constraints is thus necessary to approach a better solution. Soft orthogonal constraints (15, 18, 19) on the fragment spectra \( S \) are particularly effective for narrowing down the solution candidates, accounting for low yet non-zero possibilities that different fragments occupy common channels (full formulation in Methods section). NMF is called “unique” when \( Q \) is limited to the identity or permutation matrices (17). Importantly, the first NMF, \( X \approx AS \), is not unique because this is a spectral interpretation with no correct answer, whereas the second NMF, \( A \approx CB \), should be unique to determine the polymer fraction \( C \) with a single correct answer (20). The second NMF uniqueness can be assured by minimizing the volume of the simplex, which is spanned by row-vectors of \( B \) and contains all the datapoints, i.e., row-vectors of \( A \) (16, 21, 22). The connection between the two non-unique and unique NMFs is key to the RQMS algorithm. For robustly outputting \( (C, B) \) from any “good enough” \( A \), which is a non-unique solution of the first NMF, the factorization residual of the second NMF is evaluated using the Riemann metrics (23, 24), considering the non-orthogonality of \( S \) (fig. S6). Component number \( M \) for the first NMF is unknown and thus statistically determined via automatic relevance determination (ARD) (15, 25), whereas \( K \) for the
second NMF should be appropriately given depending on the analytical purpose (fig. S1). Two formulations were thus separately derived, as presented in the Methods section.

**Fig. 2. Benchmark compositional analysis of ternary E/M/S films.** (A) The fraction \( C \in \mathbb{R}^{N \times K}_+ \) is depicted as a \((K-1)\)-simplex by using the sum-to-one constraints \((K=3)\). There were no datapoints on the vertices (reference-free). RQMS-inferred basis spectra \( P \in \mathbb{R}^{K \times D}_+ \) were superimposed on the observed pure E/M/S spectra having absolute intensities. (B) Latent hierarchical structure of pyrolysis-MS. FA of pure polymers \( B \in \mathbb{R}^{K \times M}_+ \) indicated that the green, orange, and blue fragment spectra were mainly generated from polymer 1, 2, and 3, respectively. The black spectra were significantly generated from more than one polymer species. Notably, the characterization of the complicated fragment spectra is unnecessary.
RQMS sequencing consists of two steps: inferring $K$–basis spectra $P$ of sequence-defined copolymers and deconvoluting targeted spectrum $X_t$ into $P$ (Fig. 1B). For triad sequencing, binary and $J$–multi ($J \geq 3$) monomer systems require the basis sets of $K = 5$ and $K = J C_3 + 3 J C_2 + J C_1$, respectively. The three terms correspond to ternary, binary, and unary triads. As $J$ of $K$-bases are always easily synthesizable and, thus, referenceable homopolymers, the inference of the other bases should be reliable even if the dataset is biased, as demonstrated in the benchmark test. For sequencing, a “biased” dataset indicates that at least one of the $K$-polyads does not occupy a sufficiently high fraction in any sample.

The reliability of RQMS sequencing now depends on the validity of the linear combination model (Fig. 1A). We began with M/S triad sequencing because the alternating copolymer (MS) is exceptionally synthesizable (26), which allowed us to compare the inferred and measured (MS)$_i$ spectra. A dataset of 30 M/S random copolymers, quickly prepared via free-radical copolymerization (Data S3), was subjected to RQMS, outputting five–basis spectra attributable to (MMM)$_i$, (MMS)$_i$, (MS)$_i$, (SSM)$_i$, and (SSS)$_i$ based on the triad peaks (fig. S7). Furthermore, the inferred and observed (MS)$_i$ spectra showed good consistency. No explicit instructions specifically oriented to sequencing were implemented in the RQMS algorithm; thus, these results elucidated that sequencing can be simplified to compositional analysis and verified the underlying linear combination model.

M/S/poly(butyl acrylate) (B) ternary triad sequencing, based on a dataset of 4 terpolymers and 80 binary copolymers (Data S4), output all 13 rational basis spectra, including ternary-alternating (MSB)$_i$ (fig. S8). In contrast to NMR sequencing, which is limited to $J = 2$, RQMS sequencing has no limitations for $J$.

S/B binary pentad sequencing necessitated a more careful dataset design based on the monomer-reactivity ratio (fig. S9). Based on the 80 S/B binary copolymers (Data S5), RQMS output nine basis spectra with reasonable pentad peaks (Fig. 3A). The output tetrad/triad distributions were also appropriate; e.g., (SB)$_i$ had a single tetrad of BSBS (479 m/z) and two triads of BSB (361 m/z) and SBS (337 m/z).

After the inference of basis spectra $P = BS$, spectral deconvolution is conducted by sequentially projecting target spectrum $X_t$ onto the subspaces non-negatively spanned by rows of $S$ and $B$ (see Methods). These subspaces are peculiar to monomer combinations and invariant in polymerization systems, because unknown fragments originating from the system differences are filtered out when projected onto $S$, allowing semi-real-time direct sequencing from the polymerization solution without any chemical purification. By coupling with living polymerization (2, 27), varying polyad distributions along the main chain were monitored (Figs. 3B–C). We selected S/B copolymers as targets, since their B-centered triad fractions can be obtained via NMR as well by decoupling their tacticity (12) (fig. S10). The triad fractions (Fig. 3C) converted from the RQMS pentad fractions (Fig. 3B) were consistent with the NMR observations and theoretical predictions by Alfrey–Mayo first-order Markov terminal model based on the monomer-reactivity ratio ($r_S, r_B$) = (0.70, 0.17) (28) (also see the Methods section).
Fig. 3. S/B pentad sequencing. (A) Nine inferred basis spectra. (B–C) Sequence modulation along the main chain of living copolymers (S/B monomer feed ratio: 1/1). SSSSS, SSSBB, BBBSS, BBBB were negligible (<1%) and not shown. (B) The B-rich pentad fractions of BBBBS and BBSBS increased as the copolymerization progressed, reflecting the lower monomer reactivity of B as compared to S. The RQMS pentad fractions were downgraded to the triad fractions (C) to make the analytical results comparable to the NMR results and Alfrey–Mayo prediction.
RQMS allows quantitative analysis for inherently non-quantitative MS without necessitating any prior knowledge of the constituent polymers and spectral characterization, making it an effective tool in polymer science. RQMS sequencing would play a fundamental role for optimizing sequence distribution via sequence–property correlation studies and sequence-controlled polymerization with the cooperation of material informatics (29) and reinforce-learning (9).

References and Notes


Acknowledgments:

**Funding:** Core Research for Evolutional Science and Technology program of Japan Science and Technology Agency under Grant JPMJCR19J3 (MN).

**Author contributions:** YH conceived the research, prepared the samples, developed the software, analyzed the data, and wrote the manuscript. SU conducted the pyrolysis-MS and NMR measurements. MN supervised the research.

**Competing interests:** YH and MN are owners of patent applications on RQMS and RQMS sequencing.

**Data and materials availability:** All data are available in the main text or the supplementary materials.

Supplementary Materials

Materials and Methods
- Materials
- A synthetic procedure of random copolymers via free-radical copolymerization
- A synthetic procedure of B/S random copolymers via living radical copolymerization
- Sample preparation for the benchmark test of the E/M/S ternary polymer film
- Pyrolysis-MS measurements
- Mathematical notations
- Formatting spectra
- Derivation of the first NMF
- Derivation of the canonical correlation analysis filter (CCA-filter)
- Non-negative least square fitting
- Derivation of the second NMF
- Sequential projection of a target pyrolysis-MS spectrum onto the learned subspaces
- Prediction of sequence distribution from the monomer-reactivity ratio
- Determination of the S/B triad fraction via $^1$H NMR
- Downgrading the S/B pentad fraction into the B-centered triad fraction

Supplementary Text
- Definition of “Sequencing”
- Definitions of “Fragments” and “Polyads”

Figs. S1 to S10
Tables S1 to S2

References only cited in Supplementary Materials (30-45)

Data S1-S5